AN ANALYSIS OF THE EFFECTS OF DATA CORRUPTION ON TEXT RETRIEVAL PERFORMANCE

Stephen Smith
Craig Stanfill
Thinking Machines Corporation
245 First Street
Cambridge, MA 02142–1214
Phone: 617–876–1111
Electronic mail: smith@think.com, craig@think.com

December 14, 1988
DR90-1

ABSTRACT

Text retrieval systems are increasingly being used to search data generated by Optical Character Recognition (OCR) devices. Conservative estimates are that many government and civilian organizations have hundreds of gigabytes of paper data that will eventually be converted to computer readable form. Since such data is likely to be imperfectly transcribed by today's OCR systems it is important to determine how well current text retrieval techniques perform under varying degrees of text corruption. This paper presents research that shows that both relevance feedback and boolean search techniques display no significant degradation when confronted with text that has been corrupted at rates typical of current OCR devices. Furthermore, relevance feedback techniques can maintain average recall rates in excess of 90% on text in which 50% of the words have been corrupted.

MOTIVATION

Large amounts of textual data reside solely in hardcopy form. Current systems that convert such data from hardcopy to a computer readable form are not perfect and make mistakes in identifying characters from 1% to 3% of the time depending on the characteristics of the original hardcopy. As is pointed out by Kahan et al. [KAHAN 87] these rates are too high for most practical applications. Consider that: "a 95% {correct} rate corresponds to 150 errors per page, nearly three per line." [KAHAN 87]. These rates can appear to be even worse when viewed in terms of word corruption where the rates are typically five times worse than the character corruption rate. Because of these high error rates, practical applications of OCR systems often require a second pass of human directed correction before the data can be used.

It is the hypothesis of this paper that standard text retrieval techniques can perform quite well with such corrupted data, and in fact a fully successful system could be created whereby the corrupt OCR transcribed data was used for search purposes and the original scanned image could be stored on CD–ROM and viewed by the user. With such a system in mind a series of experiments have been carried out to determine the performance of a relevance feedback system [SALTON 71] [VAN RIJSBERGEN 79] and a simple boolean system in the face of varying amounts of data corruption.

DESCRIPTION

The text search engine used for these experiments is a modified version of the Connection Machine Document Retrieval System described in [STANFILL 86]. This system employs a compressed text data structure that is stored and searched in parallel on the Connection Machine [HILLIS 85]. Each element or *signature* of the data structure stores a chunk of thirty consecutive words that has been compressed into a one kilobit hash table. These signatures are stored one per processor on the Connection Machine and are searched in parallel. Using this data structure the CM2 Connection Machine System can currently search a database of up to two gigabytes.

The Connection Machine Document Retrieval System can perform both relevance feedback and simple boolean searches. A relevance feedback search is accomplished by first extracting content bearing words from documents that have been noted as "good" by the user. Weights are then assigned to these words based on their frequency in the database as a whole and their frequency with respect to the documents that have been marked as "good". The words and their weights are then broadcast to the Connection Machine and all signatures that contain the broadcast word have their score incremented by the weight of the query word. Proximity of important words in neighboring signatures is noted after all the words of a given marked document are broadcast. The weights of individual signatures are increased by the weights of their neighbors. The maximum score for all sections of each document is then stored in the first signature of each document. Each marked document or section of a document is broadcast in this way and on each broadcast the score generated by the current query document is compared with a running maximum score of all previous query documents and the maximum score is retained.

This algorithm can be viewed as assigning a score to a given document based on the highest scoring section of the document rather than on some cumulative measure of score over the entire document. From the user's point of view this technique allows one to zero in on the most important part of a long document.

Boolean search is implemented on the system by sequentially broadcasting each word in a boolean query, noting the signatures in which it is contained, and then performing the correct boolean combination operation between the currently broadcast word and the previously broadcast words of the query. The boolean query constructors for this system consisted of AND, OR, and NOT. No proximity, stemming, synonym generation or special text fields (eg. date or title) were allowed. An example of a boolean query would be:
(:and "bush" "dukakis" (:or "election" "campaign"))

The experiments described in this paper were performed on a 30 megabyte database of 3500 Wall Street Journal articles from January 1988. A sample of the text is shown below:

> Bethlehem Steel Profit Doubled In 4th Quarter
> By J. Ernest Beazley
> Staff Reporter of The Wall Street Journal
> 01/28/88
>
> Bethlehem Steel Corp., aided by robust prices and an upturn in capital goods markets, reported more than double fourth quarter earnings.
> The nation's No. 3 steelmaker turned in its fifth quarterly profit in a row, earning $71.5 million, or $1.07 a share, compared with $34.2 million, or 55 cents a share, a year earlier. Sales jumped 18% to $1.21 billion from $1.02 billion.

EXPERIMENTAL METHOD

To simulate a wide range of OCR performance the Wall Street Journal data was corrupted to 13 different degrees, including complete corruption. Several relevance feedback queries and boolean queries of different complexities were then run on these thirteen corrupt "datasets" to calculate the degradation of each search technique with increasing data corruption. The Connection Machine Document Retrieval System ran on a 16384 processor CM1 Connection Machine with a Symbolics series 3600 Lisp Machine serving as a front end. The CM1 memory size was 8 megabytes and the combined size of all datasets was one quarter gigabyte.

The text data for the experiment was corrupted by changing random characters to the @ character with some preset probability. Only spaces and newline characters were not corrupted. This corruption was performed on the Connection Machine by loading one character per processor and then generating a random number, R (0 <= R < N), in parallel. Processors whose random number was zero mutated their character to the @ character to produce a corruption rate of 1 in N. The values of N in this experiment were infinity, 1000, 500, 100, 75, 50, 30, 20, 10, 7, 5, 3, 2, 1. These empirically correspond to word corruption rates of: 0%, 0.6%, 1%, 4.8%, 6.5%, 9%, 15.3%, 21.8%, 38.5%, 49.7%, 61.2%, 78.7%, 90.4%, and 100%.

An example sentence at various levels of corruption is presented below:

Program Trade Limit Extended By Big Board
Program Trad@ Limit Extended By Big Board
Prog@am Trade Limit E@@en@ed By Big Board
@r@@@@m @r@d@ @@m@@ E@@@n@ed @@ Big @o@@d
@@@@@@@ @@@@@ @@@@@ @@@@@@@@ @@ @@@ @@@@@

Reference sets (collections of documents defined to have 100% precision and 100% recall) were created by first carefully defining a desired topic and by then extensively searching for that topic with both relevance feedback and boolean methods. Three reference sets were built and are discussed in the research results. They are described below:

Program Trade – A fifteen document cluster concerning specific restrictions imposed or considered to curb program trading's deleterious effect on the stock market.
Scientific Systems – A cluster of four documents concerning the takeover of the Scientific Systems company.
President Chiang – A cluster of twelve documents concerning the death of Taiwan's president Chiang and its affect on the Taiwanese government.

Boolean queries were constructed by hand to optimize recall and precision for each reference set. These queries and several single word queries were easily evaluated with standard recall and precision methods. The relevance feedback queries, however, were more difficult to evaluate since documents are returned from a relevance feedback search in ranked order rather than as an unordered hit list. To make comparisons between boolean and relevance feedback systems in terms of recall and precision some number of documents, W, was arbitrarily chosen to be returned from each query. For these experiments recall and precision were calculated with W equal to twice the size of the reference set. Thus the maximum precision could be 50%.

To enable better evaluation of subtle changes in the search effectiveness a measure called *drift* was introduced (see similar measures in [WILENSKY 89]). *Drift* attempts to measure how far a given ranked search has "drifted" away from a perfect search. The need for such a measure can be recognized by comparing the following two simulated searches. The reference set, in the example, contains five documents. Documents that are included in the reference set are denoted by i and those excluded from the reference set are denoted by x. Here are two possible retrieval cases:

Position:	1	2	3	4	5	6	7	8	9	10	. . .
Perfect Retrieval:	i	i	i	i	i	x	x	x	x	x	. . .
Example Retrieval:	i	i	x	x	i	x	x	x	i	i	. . .

If recall and precision measures are performed on this data as specified above we note that both searches have identical scores of 100% and 50% respectively. There is, however, a big difference between the quality of the two searches from the user's point of view. The drift measure is designed to reflect this difference. Drift is calculated by summing the difference between the optimal position for a document and its actual position over all included documents. Thus in the example retrieval above the third included document is in position five. It should be in position three (as in the perfect retrieval) so the drift for this document is two. Likewise the fourth included document is found at position nine but should be found at position four. Its drift is five. The drift for the example retrieval itself is the sum of the drifts of its constituent documents. In this case it is 2 + 5 + 5 = 12. The benefits of using the drift measure can be seen by comparing figure 4 and figure 3. The drift graph indicates significant changes in search effectiveness at low corruption rates that are not reflected by either the precision or recall measure.

SIMULATION DESIGN

To test the effects of different queries on the corrupted datasets each query was constructed on the non–corrupted dataset and then applied to each of the corrupted data–sets. Boolean queries were retained as the original combination of words, conjunctions, and disjunctions. Relevance feedback queries were saved as a list of seed words, good documents and good paragraphs. Queries built from good documents and paragraphs had all corrupted words stripped out before the query was broadcast.

Since the intent of this project was to evaluate the performance of a system designed for a human user the entire saved relevance feedback query could not be used as the initial query. A human user would build such a search incrementally and would not have the ability to mark documents as "good" until they had been retrieved. Such user interaction was simulated in the system by first searching on the seed words of the saved query and then marking relevant documents and paragraphs only as they appeared within the first W documents of the returned search. W is again twice the size of the reference set. This cycle was repeated until there was no further improvement in the search.

RESULTS

The results of these experiments are displayed in figure 1. This graph shows recall and precision figures for increasingly corrupted data as measured in the percentage of characters that have been corrupted. The three lines correspond to single word searches, boolean searches, and relevance feedback searches. The boolean and relevance feedback searches are averaged over the first three reference sets described above. Searches for all documents containing single words were also performed as a baseline for comparison. These searches are averaged over three different words of lengths 3, 7, and 9 letters. Note that the precision for the word probes is not always 100%. This is because an irrelevant document may contain a word that when corrupted will actually match the probed word. An example of this was found in searching for the word "sex" and finding a document in the 10% corruption dataset that had the word "Essex" that had been corrupted to "@@sex".

The actual datapoints for these graphs are listed below:

Corruption:		**Word:**		**Boolean;**		**Relevance;**		
Character	*Word*	*Precision*	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*	*Drift*
0%	0%	100	100	70.9	95.4	48.9	97.8	9.3
0.1%	0.6%	100	99.1	70.9	95.4	50.0	100.0	1.0
0.2%	1%	100	100	71.0	93.0	48.9	97.8	11.3
1%	4.8%	100	94.1	71.2	95.4	48.9	97.8	14.0
1.3%	6.5%	100	92.5	71.5	95.4	48.9	97.8	12.7
2%	9%	100	92.0	71.1	93.2	48.9	97.8	17.0
3.3%	15.3%	100	75.1	72.2	93.0	48.9	97.8	13.0
5%	21.8%	98.8	76.3	72.1	88.6	48.9	97.8	7.7
10%	38.5%	98.5	53.1	73.8	79.0	48.9	97.8	41.3
14%	49.7%	96.8	52.2	73.1	38.3	46.7	93.3	69.3
20%	61.2%	97.6	22.6	79.7	42.6	40.3	80.6	174.7
33%	78.7%	59.3	13.8	66.7	11.4	31.1	62.2	349.0
50%	90.4%	53.3	9.23	0	0	12.6	23.9	804.0
100%	100%	0	0	0	0	0	0	1033.3

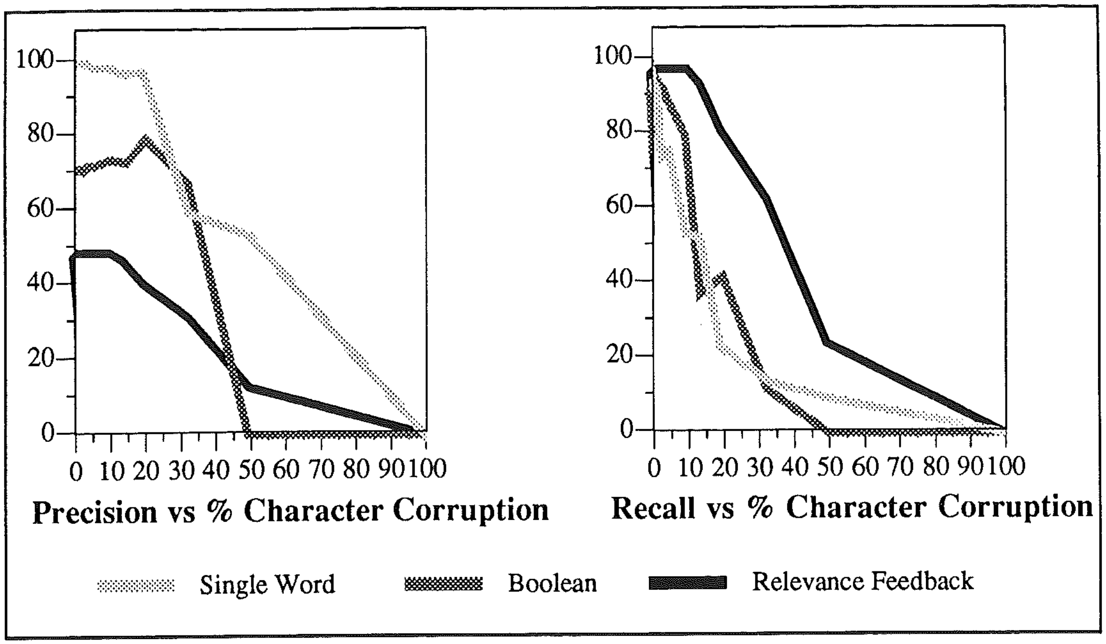


figure 1

The graphs of figure 2 contain the same data as in figure 1 but the horizontal axis is in terms of word corruption rather than character corruption. Viewing the data in terms of word corruption is more relevant to the text retrieval techniques researched here in that they are based on the assumption of an exact word to word match. Character corruption would be a more useful measurement for character level search techniques such as n–gram analysis [KIMBRELL 87].

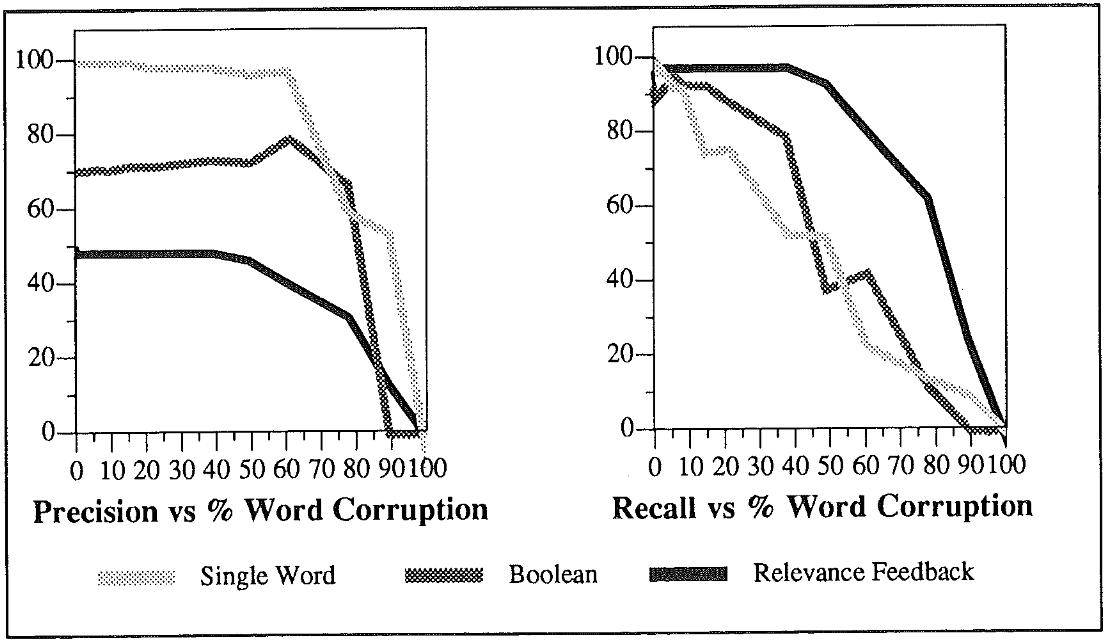
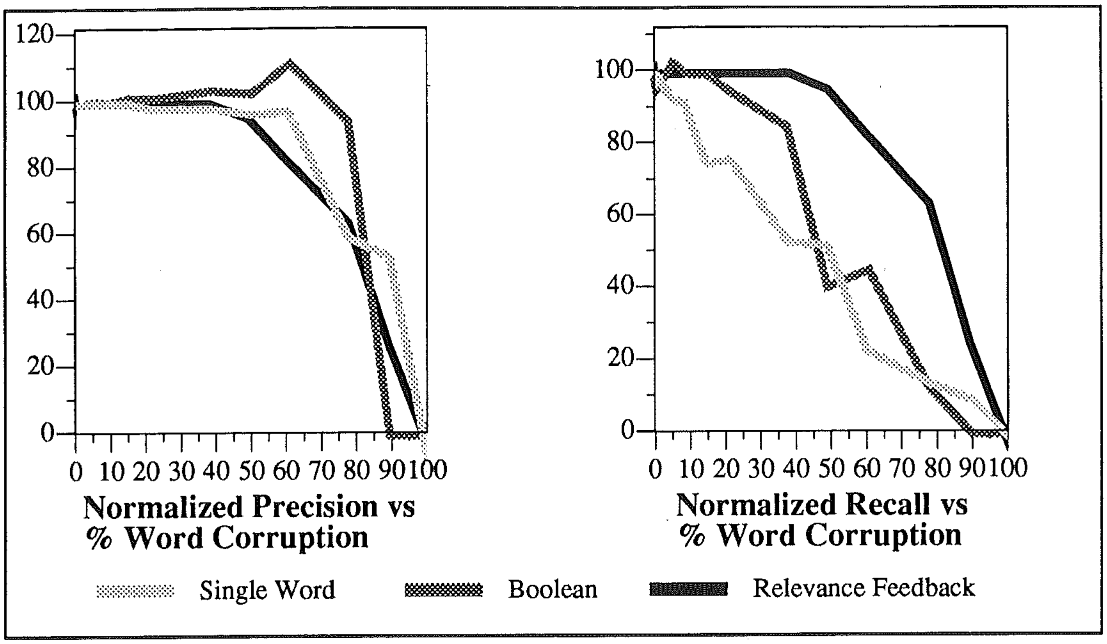


figure 2

As mentioned earlier the maximum precision attainable in this experiment by relevance feedback was 50% because of the experimental setup. The graphs in figure 3 are the same as those in figure 2 except that each data point has been normalized to the performance on the initial uncorrupted data. Relevance and precision measures begin at 100% and the graphs reflect the change in degradation rather than the absolute performance of the system.

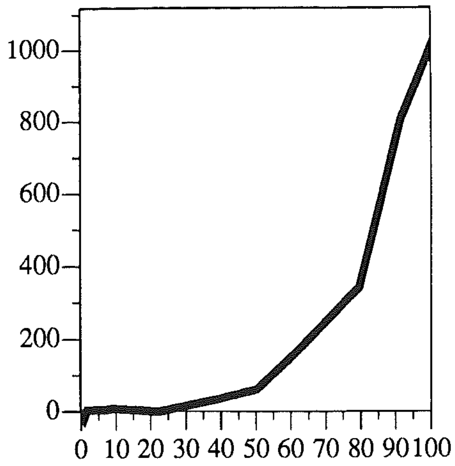


Normalized Precision vs % Word Corruption

Normalized Recall vs % Word Corruption

figure 3

The graph in figure 4 shows drift as a function of data corruption. This graph is interesting in that it gives a more fair accounting of the degradation taking place in the relevance feedback system. Where recall and precision show no change at low degrees of corruption for relevance feedback, the drift measure, more accurately, shows a consistent degradation.



Drift vs % Word Corruption

Relevance Feedback

figure 4

ANALYSIS

The most important result of these corruption experiments is that for low levels of corruption (well within the capabilities of OCR devices) there is almost no degradation of either the boolean or relevance feedback search techniques. The introduction of small quantities of noise even seems to improve the recall of the relevance feedback system. Another important observation is that relevance feedback outperforms boolean search. Surprisingly, a relevance feedback system can achieve 20% recall with 50% of the characters or 90% of the words corrupted.

Further research into this specific high corruption case showed that, for the program–trade reference set, the two relevant documents that were found were located by the single word "trading". In each of the documents there were at most 19 other words that were left uncorrupted and none of them were relevant to the query. Thus no iterative search refinement could be performed. It would be interesting to see whether the improvement of relevance feedback search over boolean search is due solely to the looseness with which relevance feedback queries are defined or whether the marking of relevant documents aids in the search refinement. This is not currently known.

It can by hypothesized that the *tighter* a search query becomes the more susceptible it will be to corrupted data. The quintessential *tight* query would be a long conjunction of words. The opposite of a *tight* query would be a *loose* query and a quintessential *loose* query would be a long disjunction of words. For a simplified case, where every document in the database contained all the words in the query, not retrieving a relevant document for a disjunctive query means that every query word in the document would have to be corrupted. For an N word query and a corruption rate of C the probability of missing a document with a disjunctive query would be: C^N. For a *tight* conjunctive query in the same database the probability of not retrieving a document would be equivalent to the probability that any one of the query words contained in the document was corrupted. To calculate this it is best to think of the problem as one minus the probability that every word in the document was not corrupted or: $1-(1-C)^N$. A few values for the probability of missing a document with a *tigh*t conjuncitve or a *loose* disjunctive query are listed below:

N	C	CONJUCNTION $1-(1-C)^N$	DISJUNCTION C^N
1	.1	.1	.1
3	.1	.27	.001
10	.1	.65	1e–10
100	.1	.99	0

With this analysis it is possible to see that any conjunction in a boolean search tightens the query. Since relevance feedback queries do not contain explicit conjunctions they can be characterized in general as looser than boolean queries and less susceptible to data corruption.

Another factor which may contribute to the success of relevance feedback is that the documents are in essence competing amongst themselves to match the query rather than trying to match the query in an absolute sense. The advantage is that, though, the scores of relevant documents may decrease as the data is corrupted, they may still be the best matches in the database since all the other document match scores were decreased by a similar amount. For instance if every document in a database is corrupted equally then the strength of match between a query and a relevant document will be lower than in a non–corrupted database. Similarly all irrelevant documents are equally likely to be corrupted and their match strength should decrease also. Thus the relevant documents would be returned ranked in the proper order for the search even though their overall scores have been decreased.

The number of words in a loose query may also play a role in performance under data corruption as can be seen from the analysis done above. This may contribute to the succes of relevance feedback over boolean systems since relevance feedback queries are much larger than the typical boolean query. A full boolean query for the program–trade reference set, for example, contains 42 words. A full relevance feedback query for that reference set contains 1344 different words and 2636 total words.

LIMITATIONS OF RESEARCH

There are several limitations to this research one of the most important of which is that these studies have been performed with simulated OCR data. It is possible that a system would perform differently when presented with actual OCR data. For example the assumption that characters will be "rejected" rather than falsely detected is not necessarily consistent with OCR performance. Most manufacturers, however, strive for rejections rather than false detections [KAHAN 87] and can probably accommodate a greater percentage of rejections for correspondingly lower overall performance. This should be acceptable for the boolean and relevance feedback systems since they seem to be able to perform well with corruption rates substantially greater than those produced by current OCR systems.

The importance of such performance in retrieval systems might also be questioned since OCR research is progressing so rapidly that it may be only a matter of years before 100% recognition rates are achieved. This may be true but in addition to the clean machine printed text currently being worked on there will always be more difficult OCR problem such as human handwriting [AHMED 87][SABOURIN 86] that will not meet with 100% recognition rates.

A fair criticism of these experiments is that only a limited number of reference sets were run and that the boolean system in use did not employ such techniques as stemming and proximity. Though more reference sets will be run, there was little variance in degradation curves for the datasets that were investigated. Thus it would be unlikely to see a large change in performance with further testing. Though the boolean syntax that was used for these experiments did not include stemming a strong attempt was made to produce morphological variants of words in the queries and to combine them with the OR constructor. It is believed that stemming and word proximity information would have increased the initial performance of the boolean system but would have not have compensated for the problems of conjunctions used in boolean constructs. Thus the rate of degradation would probably have been about the same.

FUTURE DIRECTIONS

The research presented in this paper shows that boolean and relevance feedback search techniques perform well with corrupted data. There is not yet, however, a theory of performance degradation, or a theoretical explanation of the superiority of a relevance feedback search over a boolean search. An information level analysis of these results should be the main focus for future research on this problem.

ACKNOWLEDGMENTS

Thanks are extended to the Dow Jones Corporation for the use of their data, and to Brij Masand, Alex Singer, Dave Waltz and Uri Wilensky for their suggestions and insights.

REFERENCES

[AHMED 87] Pervez Ahmed and C. Y. Suen, “Computer Recognition of Totally Unconstrained Hadwritten Zip Codes” International Journal of Pattern Recognition and Artificial Intelligence, Vol. 1 No. 1 (1987) pp. 1–15.

[KAHAN 87] Simon Kahan, Theo Pavlidis, and Henry Baird, “On the Recognition of Printed Chracters of Any Font and Size” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI–9, No. 2. March, 1987; pp. 274–287.

[HILLIS 85] Daniel Hillis, The Connection Machine, MIT Press, Cambridge MA, 1985.

[KIMBRELL 87] Roy E. Kimbrell, “Text: Storage and Retrieval Using N–Grams”, Private Manuscript.

[SABOURIN 86] Robert Sabourin and Rejean Plamondon, “Preprocessing of Handwritten Signatures from Image Gradient Analysis”, Proceedings of the IEEE International Conference on Pattern Recognition, 1986, pp. 576–579.

[SALTON 71] Gerard Salton, The SMART Retrieval System – Experiment in Automatic Document Processing, Prentice–Hall, Englewood Cliffs, N.J., 1971.

[STANFILL 86] Craig Stanfill and Brewster Kahle, “Parallel Free Text Search on the Connection Machine System”, Communications of the ACM, Vol. 29 No. 12, December, 1986, pp. 1229–1239.

[VAN RIJSBERGEN 79] C. J. van Rijsbergen, Information Retrieval, Second Edition, Butterworths, London, 1979.

[WILENSKY 89] Uriel Wilensky, Personal Communications on Document Retrieval Evaluation Project; Thinking Machines Corporation, 1989.